ComiRNet User Guide (v. 1.2)

- Introduction
- About ComiRNet
  - Method
- Query Functions
  - Search Interactions
  - Search Biclusters
- Tutorial

Introduction

This document provides an overview of ComiRNet content and utilities. This is not a comprehensive guide, but should provide users with enough information to properly browse the database and use its principal tools for data analysis.

Please read through it and contact us at *gianvito.pio_AT_uniba.it* with any comments or questions.

About ComiRNet

ComiRNet (Co-clustered miRNA Regulatory Networks) is a database specifically designed to provide biologists and clinicians with user-friendly and effective tools for the study of miRNAs.

The database stores automatically-mined and non-redundant data of miRNA-gene target interactions (MTIs) and miRNA-gene regulatory networks (MGRNs) in the form of biclusters. Data are produced by exploiting miRNAs target predictions from 10 different prediction databases stored in mirDIP and validated MTIs extracted from miRTarBase.

Based on the principles of the ComiRNet approach, genes in a bicluster are likely to function together as a network and miRNAs in the same bicluster are likely to cooperatively target groups of networked genes.

The use of computational predictions in place of only experimentally validated interactions offers the possibility to detect single interactions and regulatory modules that would be otherwise impossible to reconstruct by considering only experimentally validated interactions, which are strictly dependent on the cell type and experimental conditions used. This paves the way to the systematic use of ComiRNet for:

1. a comprehensive analysis of cooperative targeting of miRNAs of interest (Figure 1a);
2. the discovery of unknown miRNA and gene functions, on the basis of the ComiRNet biclustering (Figure 1b);

3. the discovery of unknown miRNA targets which could be worth to be experimentally validated. This possibility is due to the ComiRNet ability to associate objects that are apparently not related (Figure 1c).

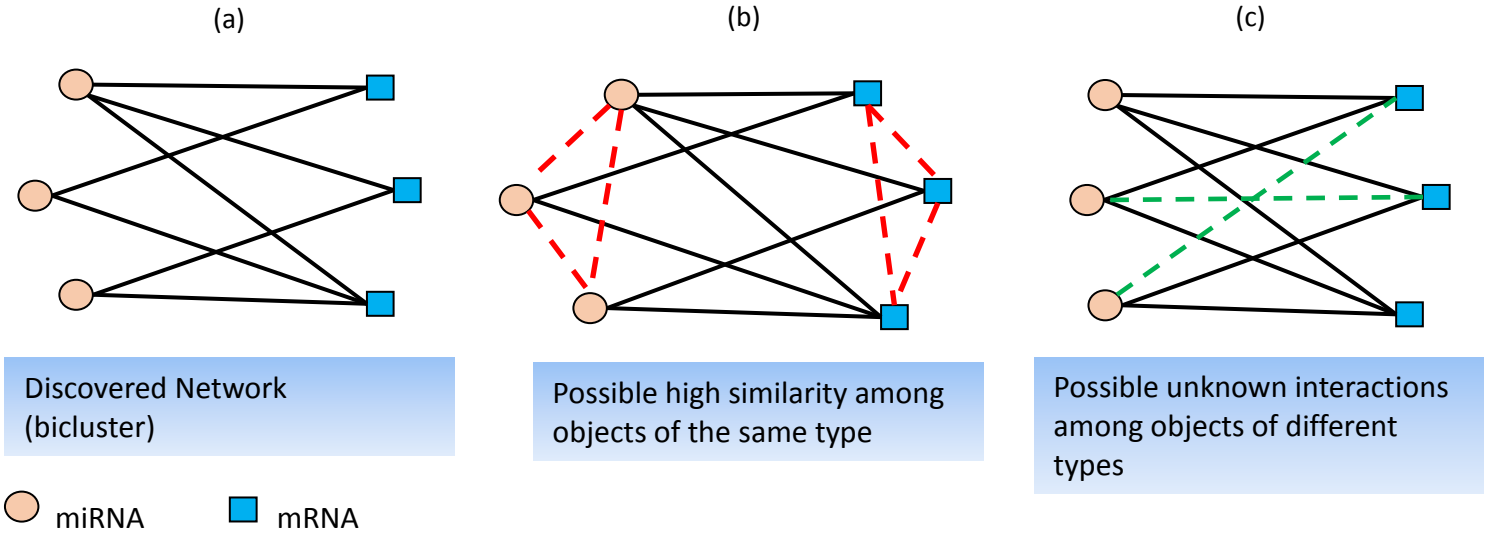

Figure 1. a) Biclusters extracted by ComiRNet suggest interaction networks between objects of different types (i.e., miRNA and mRNA), black edges indicate interactions among miRNAs and target genes in the bicluster; b) red dashes edges underline the putative functional similarity, among object of the same type, suggested by the ComiRNet biclustering; c) green dashes edges indicate putative unknown functional interactions between miRNAs and genes suggested by the ComiRNet biclustering.

Method

ComiRNet is based on a two-stepped computational approach (see Figure 1). In the first step, a semi-supervised ensemble-based classifier (see [1] in Publications) is learned from both experimentally validated interactions (positively labelled examples) and miRNA gene target predictions (MTIs) returned from several prediction algorithms (unlabelled examples). This classifier acts as a meta-classifier of unlabelled examples. As a result of the first step, a unique (meta-)prediction score is available for all possible interactions. In the second step, these prediction scores are used to identify miRNA-gene regulatory networks (MGRNs) through the biclustering algorithm HOCCLUS2 (see [2] in Publications).

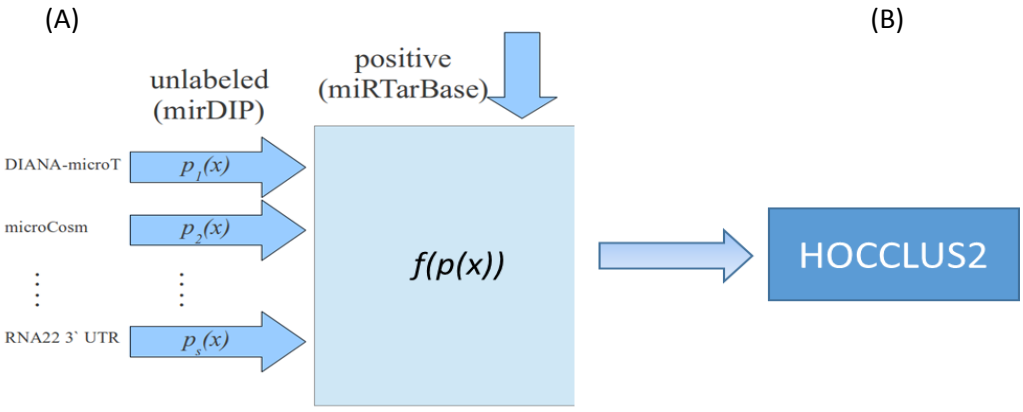

Figure 2. ComiRNet computational approach.

Step (A)

The semi-supervised ensemble-based classifier learns to combine predictions, referred to 3'-UTR of genes targeted by miRNAs, extracted from: *DIANA-microT, micro-Cosm, miRanda, picTar 4-way* and *picTar 5-way, PITA All Targets* and *PITA Top Targets, TargetScan Conserved* and *TargetScan Non-Conserved, RNA22 3' UTR.*

The algorithm is three-stepped:

1. Train an SVM classifier which outputs the probability of an instance to be labelled
2. Assign a weight to each instance on the basis of its probability of being a labelled instance
3. Train a SVM classifier which takes into account instance weights and outputs the probability of an instance to be positive.

Both steps 1) and 3) are performed by resorting to an ensemble-based learning approach. In particular. K subsets of instances are identified, consisting of the whole set of positive instances (i.e., experimentally validated interactions) and a subset of the unlabelled instances (i.e., predicted interactions), randomly sampled with replacement.

Step (B)

The biclustering algorithm HOCCLUS2 (**Hierarchical and Overlapping Co-CLUStering 2)** exploits the set of non-redundant interactions predicted by the semi-supervised ensemble-based classifier, with the associated probabilities, to identify overlapping and hierarchically organized biclusters, each one representing putative MGRNs.

HOCCLUS2 consists of three steps:

1. **Extraction of a set of non-hierarchically organized biclusters in form of bicliques** (Figure 3), through an iterative bottom-up strategy. This step exploits statistical properties of data.

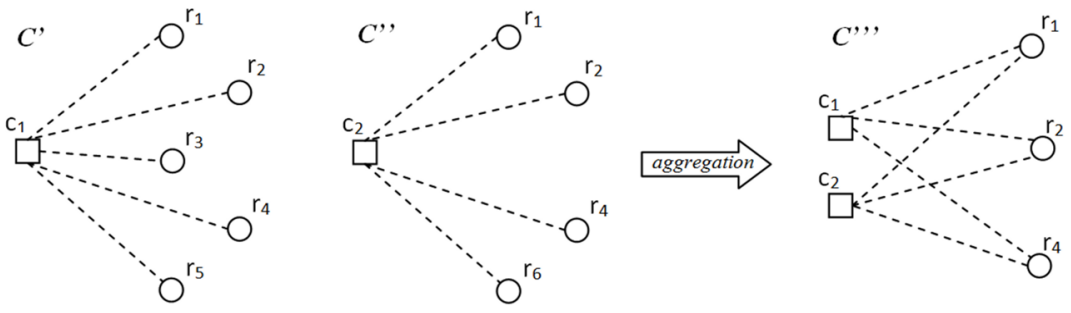

Figure 3. Identification of biclusters in form of bicliques.

2. An iterative process in which, at each iteration, two operations are performed (see Figure 4):

a. **overlap identification**, in which miRNAs or mRNAs belonging to a bicluster can be added to another bicluster, by exploiting an SVM-based classification algorithm.
b. **merging**, in which biclusters are merged when some (distance- and density-based) heuristic criteria are satisfied. Merging implicitly defines a hierarchy of biclusters (see Figure 5).

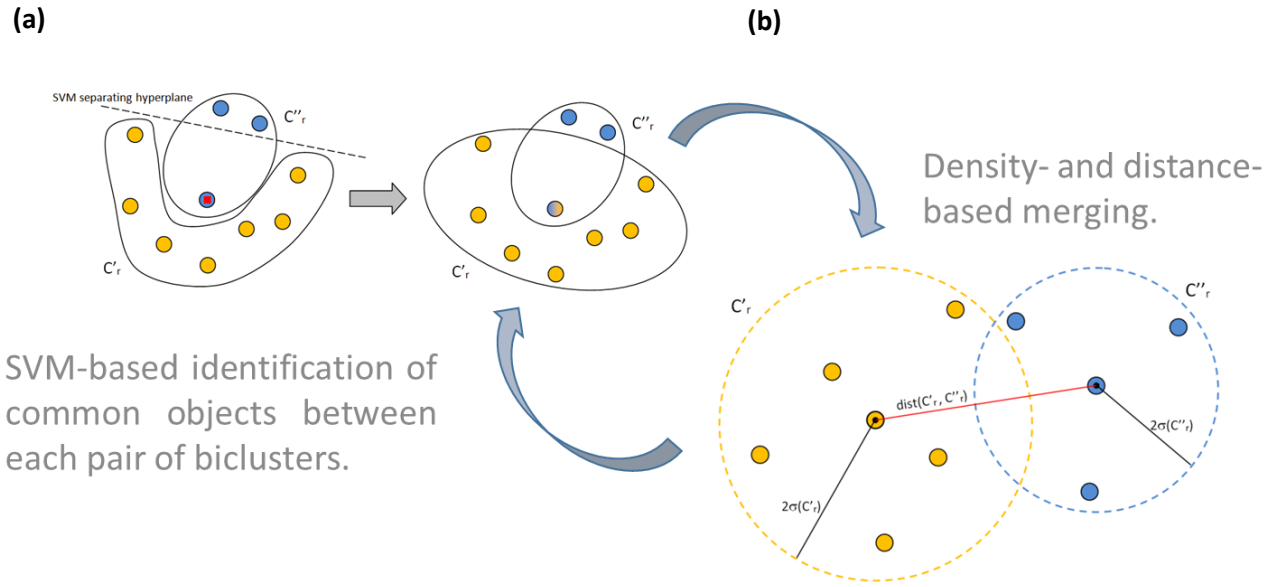

Figure 4. HOCCLUS2 –Second step of the algorithm execution: Overlap identification and merging of biclusters. The stopping criterion is based on a cohesiveness threshold.

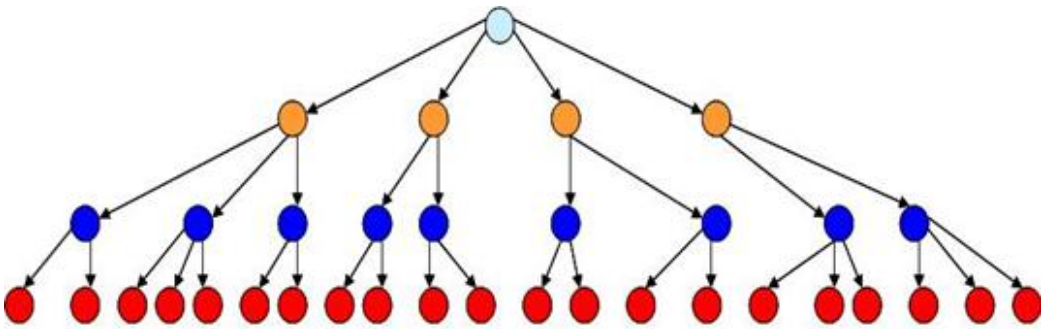

Figure 5. HOCCLUS2 – Hierarchies of overlapping biclusters. The hierarchical structure of biclusters, as provided by HOCCLUS2, helps to detect multiple alternative co-targeting of different miRNAs on specific groups of genes.

3. **Ranking of the extracted biclusters (Figure 6).** Ranking is based on the p-value obtained by Student's t-test through which we compare the average intra-bicluster similarity to the average inter-bicluster similarity, among miRNA target genes.

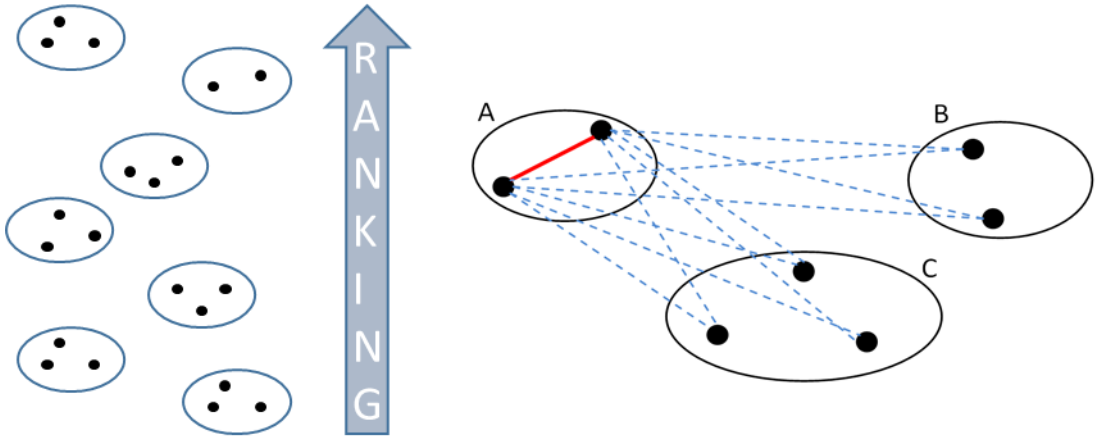

Figure 6. HOCCLUS2 – Third step of the algorithm: Ranking of biclusters. Red edges represent intra-bicluster similarities; blue dashed edges represent inter-bicluster similarities. The similarities between miRNA targets (belonging to the same and to different biclusters, respectively) are pairwise computed, according to the simGIC similarity, on the gene classification provided in Gene Ontology.

Query Functions

ComiRNet provides two main modules for querying the database, that are Search Interactions and Search Biclusters. Each module is equipped with a web interface for the retrieval and visualization of data. Several filtering criteria can be used to refine the query to satisfy specific user needs.

Search Interactions

The Search Interaction module allows users to extract MTIs on the 3'UTR of all known human genes. Currently ComiRNet stores about 5 million predicted interactions between 934 human miRNAs and 30,875 gene transcripts (mRNAs). Results are not redundant and are shown with the score (i.e. probability) identified by the approach described in Method - Step (A).

Details of the available query options are shown in Figure 7. The output consists of an interactive table visible (see the bottom of the figure). Numbered boxes help to underline, step by step, all the available options and filters that can be used to refine the query and to export the results.

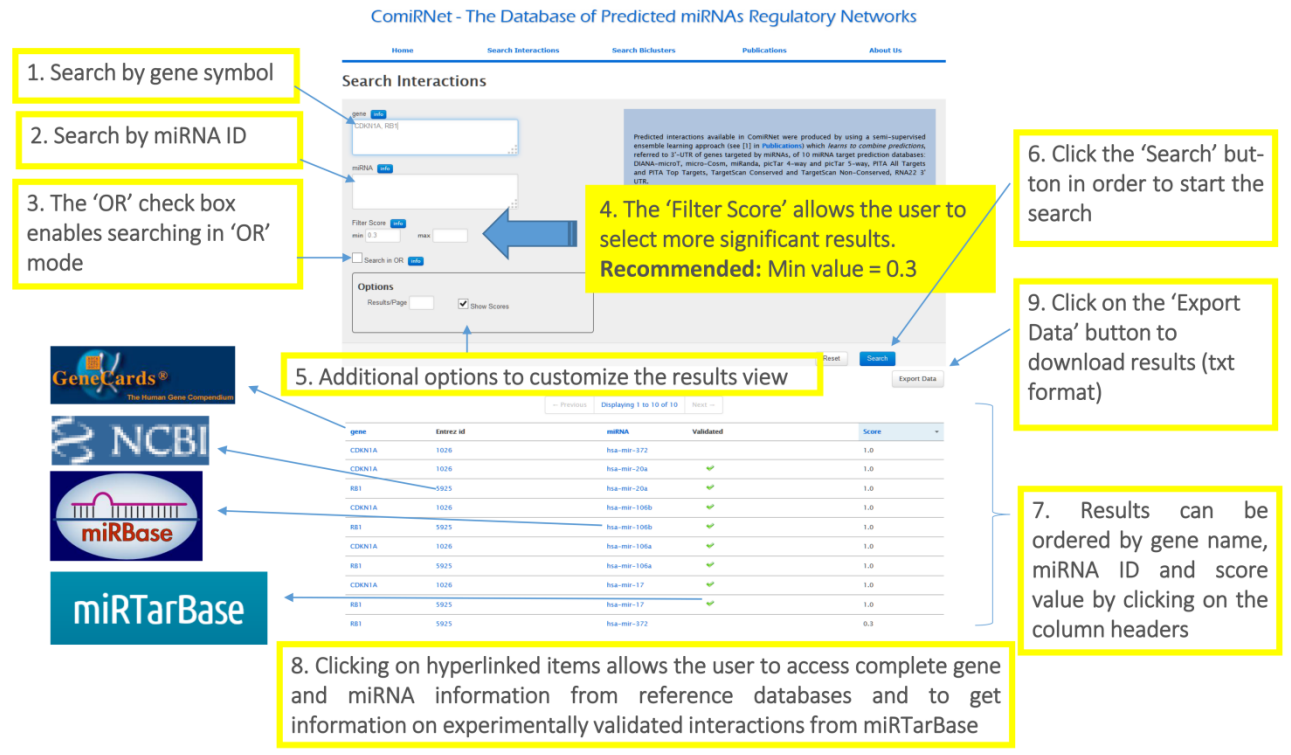
ComiRNet - The Database of Predicted miRNAs Regulatory Networks
Search Interactions
1. Search by gene symbol
2. Search by miRNA ID
3. The 'OR' check box enables searching in 'OR' mode
4. The 'Filter Score' allows the user to select more significant results. Recommended: Min value = 0.3
5. Additional options to customize the results view
6. Click the 'Search' button in order to start the search
7. Results can be ordered by gene name, miRNA ID and score value by clicking on the column headers
8. Clicking on hyperlinked items allows the user to access complete gene and miRNA information from reference databases and to get information on experimentally validated interactions from miRTarBase
9. Click on the 'Export Data' button to download results (txt format)
GeneCards
NCBI
miRBase
miRTarBase

Figure 7. Details on the options available in the Search Interaction module.

Boxes 1-2-3. MTIs can be queried by specifying one or more search items separated by commas. Gens have to be specified by using Gene official symbols (e.g. CDKN1A), whereas miRNAs have to specified by using miRNA identifiers (e.g. hsa-mir-17) (**boxes 1-2**). The system searches with the "AND" condition by default. As an alternative, the user can perform the query by enabling the "OR" condition check box (**box 3**).

Boxes 4-5-6. A filter on the interaction score in the interval [0,1] (**box 4**) allows users to perform the query at different levels of stringency. We recommend to filter interactions with the scores lower then 0.2-0.3. Indeed, lower score values would return too many interactions with a low significance. Additional options are provided in the "Options" box (**box 5**) which allows users to choose how many results (i.e., MTIs) have to be shown per page, and the inclusion/exclusion of interaction scores. Finally, the search button allows the user to start the query with the specified options (**box 6**).

Boxes 7-8-9. The result table shows the list of MTIs retrieved. In particular, it shows the gene symbol, the gene's ENTREZ ID, the miRNA ID, a green check symbol (if the interaction is validated in miRTarBase) and the interaction score. Results can be ordered, by clicking on the column header, with respect to gene symbols, ENTREZ IDs, miRNA IDs and interaction scores (**box 7**). Complete information on genes and miRNAs are provided throughout the hyperlink to the their own entry in reference databases (GeneCards and NCBI for target genes, miRBase for miRNAs) (**box 8**). If an interaction is validated in miRTarBase, by clicking on the green check symbol the user is brought to the relevant entry in the reference database. Finally, it is possible to export and download the query results, by clicking on the "Export Data" button (**box 9**).

Search Biclusters

ComiRNet also stores MGRNs predicted by HOCCLUS2 on the basis of the identified MTIs, that can be queried through the "Search Bicluster" module.

Bicluster Properties

Each MGRN, represented as a bicluster, is characterized by several properties that help the user in the selection of the most significant MGRNs on the basis of different criteria, that are:

- **The bicluster compactness**. This value can vary in the interval [0,1] and measures the bicluster cohesiveness. The compactness of a bicluster represents the weighted percentage of direct interactions in the bicluster, normalized by the number of all the possible interactions. The higher the compactness value, the higher the probability that objects in the bicluster are involved in the same pathway or in strictly related pathways.
- **The intra-bicluster biological coherence**. This quantity is expressed by the value of two parameters, i.e. pBP and pMF, which measure the similarity of target genes in the bicluster (with respect to genes in other biclusters) on the basis of the biological process (BP) in which they are involved, or of their molecular function (MF). The lower the p-values the higher the probability that: i) genes in the bicluster are involved in the same biological process or that many of them have related molecular functions; ii) miRNAs in the bicluster work together as a regulatory module.
- **The level of the hierarchy to which the bicluster belongs to.** The lower the hierarchy level to which a bicluster belongs, the lower the number of objects in the bicluster but the higher the percentage of them with direct interactions. The biclusters compactness gives a measure of this feature. On overall, biclusters belonging to lower levels of hierarchies are the most useful to detect pathway-specific activities of miRNAs, whereas biclusters at higher levels are much more informative about inter-pathway functional correlations.

Biclusters Source

ComiRNet stores 15 different hierarchies (defined as 'Source' in the search form), obtained by varying the threshold values of two parameters of HOCCLUS2, i.e. alpha and beta.

Alpha is the minimum cohesiveness value that a bicluster must satisfy after performing a merging. The value of this parameter implicitly influences the number of the hierarchy levels and the number of biclusters at each hierarchy level (i.e. the higher the value of alpha, the lower the number of biclusters per level).

Beta is the minimum score that an interaction must have to be considered as reliable. The higher its value, the more the predicted interaction networks are reliable, but the less is their number.

Table 1 shows some statistics about the hierarchies identified by HOCCLUS2, considering the number of hierarchy levels and the number/percentage of significant biclusters (p-value < 0.05) per hierarchy, for each combination of alpha and beta thresholds.

HIERACHY			# level	# biclusters	pBP <0.05	pMF < 0.05
id	alpha	beta				
1	0.1	0.3	8	1861	576 (30.95%)	515 (27.67%)
2	0.1	0.4	8	1229	377 (30.67%)	349 (28.39%)
3	0.1	0.5	8	866	309 (35.68%)	260 (30.02%)
4	0.2	0.3	7	2172	654 (30.11%)	639 (29.41%)
5	0.2	0.4	7	1399	443 (31.66%)	408 (29.16%)
6	0.2	0.5	7	966	350 (36.23%)	287 (29.71%)
7	0.3	0.3	6	2469	755 (30.57%)	674 (27.29%)
8	0.3	0.4	6	1570	485 (30.89%)	459 (29.23%)
9	0.3	0.5	7	1181	425 (35.98%)	398 (33.70%)
10	0.4	0.3	6	3115	873 (28.02%)	787 (25.26%)
11	0.4	0.4	6	1863	608 (32.63%)	532 (28.55%)
12	0.4	0.5	7	1371	494 (36.03%)	444 (32.38%)
13	0.5	0.3	5	3415	851 (24.91%)	735 (21.52%)
14	0.5	0.4	5	2039	623 (30.55%)	541 (26.53%)
15	0.5	0.5	5	1329	453 (34.08%)	391 (29.42%)

Table 1. Some statistics about hierarchies stored in ComiRNet.

On the basis of the values of alpha and beta, the number of hierarchical levels and the number of significant biclusters per level may vary in a sensible manner. Hence, the selection of one hierarchy to analyze is fundamental for the type of results the user can get. As for a first and general exploration of the biclusters, we suggest to start from the hierarchy 1, which is the less stringent among all the hierarchies. Once the user detects the bicluster of interest, a search in hierarchies with higher values of alpha and beta parameters can help in the retrieval of more significant results.

Details of the available query options are shown in Figure 8. Numbered boxes help to underline, step by step, all the available options and filters that can be used to refine the query and to export the results

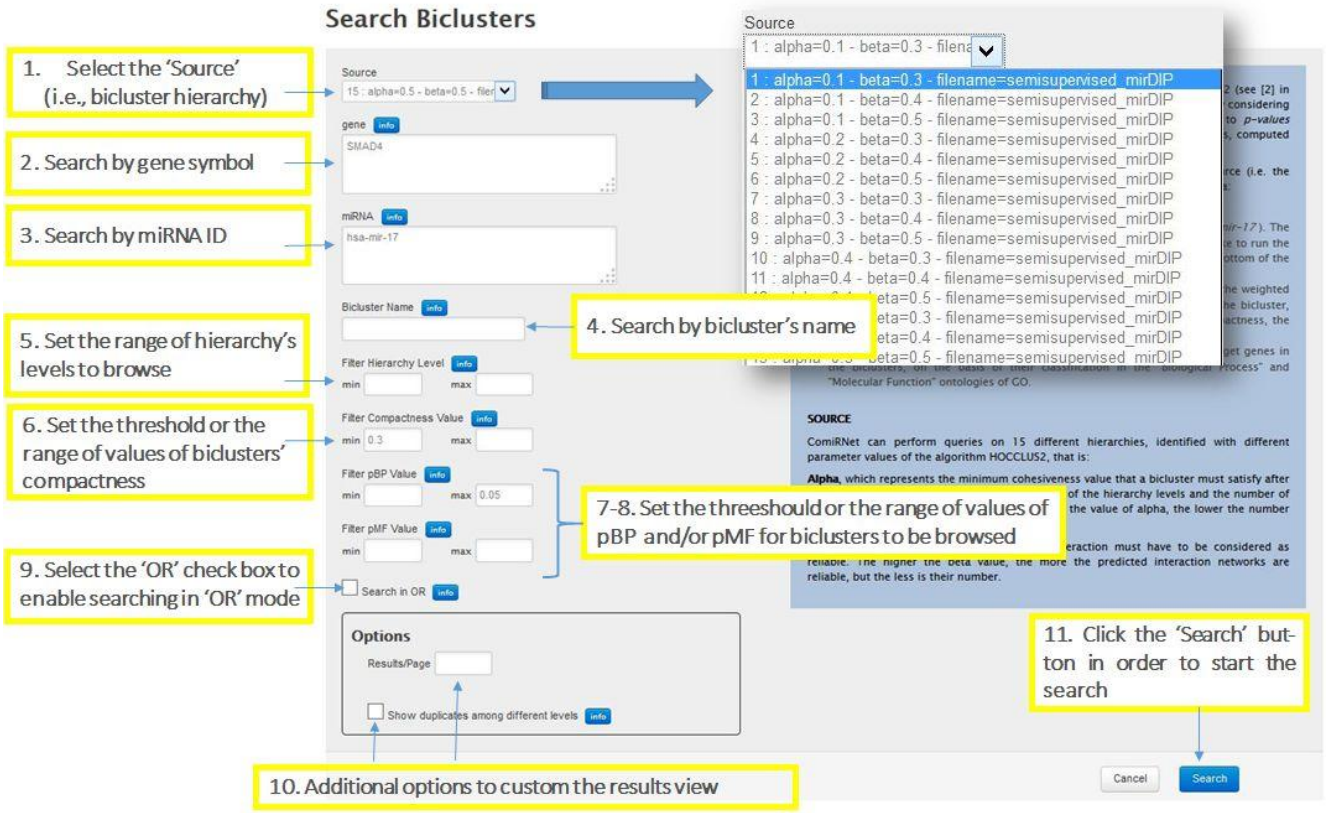
Search Biclusters
1. Select the 'Source' (i.e., bicluster hierarchy)
2. Search by gene symbol
3. Search by miRNA ID
4. Search by bicluster's name
5. Set the range of hierarchy's levels to browse
6. Set the threshold or the range of values of biclusters' compactness
7-8. Set the threeshould or the range of values of pBP and/or pMF for biclusters to be browsed
9. Select the 'OR' check box to enable searching in 'OR' mode
10. Additional options to custom the results view
11. Click the 'Search' button in order to start the search
Source
gene
SMAD4
miRNA
hsa-mir-17
Bicluster Name
Filter Hierarchy Level
Filter Compactness Value
Filter pBP Value
Filter pMF Value
Search in OR
Options
Results/Page
Show duplicates among different levels
Cancel
Search

Figure 8. Details on the options available in the Search Biclusters module. The figure shows a search in hierarchy 15 (**box 1**) using as search criteria the gene SMAD4 and the miRNA hsa-mir-17 (**boxes 2-3**). Filter applied are: biclusters compactness (**box 6**) with a min value = 0.3 and pBP ≤ 0.05.

Box 1 (source) allows the user to select the source hierarchy and it is mandatory. After selecting the desired hierarchy, two types of queries can be performed: i) the retrieval of all the biclusters in the hierarchy, or ii) the exploration of only those biclusters containing miRNA(s) and/or gene(s) of interest (**boxes 2-3**). In the latter case, similarly to the "Search Interaction" form, a list of gene symbols and/or miRNAs IDs can be provided, both as single search criterion or in combination. The system searches with the "AND" condition by default. As an alternative, the user can perform the query by enabling the "OR" condition check box (**box 9**).

The 'Bicluster name' search field (**box 4**) lets users search for a single bicluster. This feature is useful to quickly retrieve biclusters that were considered interesting in a previous analysis.

The filter on the hierarchical level (**box 5**) allows the user to select a range of hierarchical levels that the system has to consider. This filter is useful to discard useless results once the user has already analyzed the full hierarchy and identified the levels with the most interesting results. If you are using the database for the first time, we suggest do not use this filter, to avoid to discard some potentially interesting results.

The filter on the bicluster compactness (**box 6**) in the interval allows the user to run the query at different levels of stringency with respect the compactness of the biclusters that have to be selected. We recommend the use of this filter at min value 0.2-0.3 and max value not more than 0.5. Indeed, lower score values would return biclusters with too much low significance whereas score higher than 0.5 may exclude very significant results.

The filters on the pBP and pMF values (**boxes 7-8**) allow the user to select the most significant results on the basis of the p-values (pBP and pMF respectively), measuring the biological significance of the genes in the biclusters. The use of these filters with a max value of 0.05 is suggested in order to reduce the retrieval of too many results with a poor significance. In any case, we suggest to avoid using both the filters together in the same query because they do act on different biological properties of the biclusters. Indeed, the selection of the most significant biclusters on the basis of pMF can hide highly significant biclusters on the basis of pBP.

Additional options are available in the "Options" section (**box 10**), which allows users to choose how many results (i.e., biclusters per page) have to be shown and the inclusion/exclusion of biclusters' duplicate. Biclusters' duplicates can be generated at different level of the hierarchy. We suggest to keep the default option, in order to avoid redundancy in the results.

Finally, the search button allows the user to start the query with the specified options (**box 11**).

The results obtained from the query are shown in a table placed on the bottom of the search form (see Figure 9). The table includes the list of biclusters matching the search criteria and reports, for each bicluster, the hierarchy level to which it belongs, the bicluster name/identifier, the compactness value, the number of genes and miRNAs involved and the pBP and pMF values. The results in the table can be dynamically sorted according to each column and can be exported as plain text or XML file.

Figure 9. Results table of the Search biclusters module. The figure shows the results obtained by searching for biclusters satisfying the criteria specified in Figure 8.

In the rightmost column, a 'Show' button opens a new window reporting the summary of the bicluster properties (Figure 10, panel A), a dynamic graph-based visualization of the predicted miRNA-gene interactions network (Figure 10, panel B), and a comprehensive view of the bicluster hierarchy (i.e., parent and child biclusters) (Figure 10, panel C).

Figure 10. Details about a selected bicluster.

In **panel A,** the user can see bicluster properties, where the searched items (gene official symbols and/or miRNA ID) are underlined in red.

The 'Filter interactions' slider, placed on the top left side allows users to dynamically customize a threshold on the minimum score of interactions (in the interval [0, 1]) to be shown in the network graph visualization (panel B). Moving the slider from left to right, the system dynamically redraws the graph, excluding all those miRNA-gene interactions with a score below the selected threshold. Moreover, a check box allows the user to hide isolated nodes, i.e. miRNAs and genes that are not connected to any other nodes in the bicluster, according to the selected threshold. This option is particularly useful for an easier interpretation of predicted MGRNs in which a large number of miRNAs and genes is involved (e.g., biclusters belonging to high levels of the hierarchy). The application of this filter contextually modifies the list of miRNAs and genes of the bicluster, reported in the summary of bicluster properties, thus facilitating the users in keeping only those objects in the bicluster that are particularly interesting for further analysis.

Panel B contains the graph-based representation of the interaction network. Nodes represent miRNAs and target genes, whereas edges represent the miRNA-gene target interactions. By hovering the mouse pointer on a miRNA or on a gene, the user can highlight all the predicted targets or all the miRNAs targeting the gene, respectively. This allows user awareness of the impact of a single miRNA on the whole set of genes involved in the bicluster or alternatively, of which are the miRNAs, among all those in the bicluster, that co-target a specific gene. This is particularly important when the user is exploring biclusters which do not belong to the first level of the hierarchy. Indeed, in this case, biclusters do not necessarily represent fully-connected networks, and the identification of co-targeting entities becomes important. When the user hovers the mouse pointer on an edge, the system shows the predicted interaction score, enabling a quick evaluation of the reliability of a specific interaction, in the overall context of the network.

Panel C contains the "Hierarchy Browser", which allows the user to browse the hierarchy the considered bicluster belongs to, by analyzing the details of its parent and child biclusters. Similarly to the interface used to show the results of queries on MGRNs, also in this case some information about listed biclusters are provided. Detailed properties of each bicluster can be visualized clicking on the 'Show' button.

Tutorial

Click here to open a video tutorial showing dynamic functions described in this guide.